BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

E4 aus Prüf. Bescheid vom 12.03.2002

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G01N 27/26, 27/447 — A1

(11) International Publication Number: **WO 00/63683**

(43) International Publication Date: 26 October 2000 (26.10.00)

(21) International Application Number: PCT/US00/10504

(22) International Filing Date: 19 April 2000 (19.04.00)

(30) Priority Data:

60/130,238	20 April 1999 (20.04.99)	US
09/513,395	25 February 2000 (25.02.00)	US
09/513,486	25 February 2000 (25.02.00)	US
09/513,907	25 February 2000 (25.02.00)	US

(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

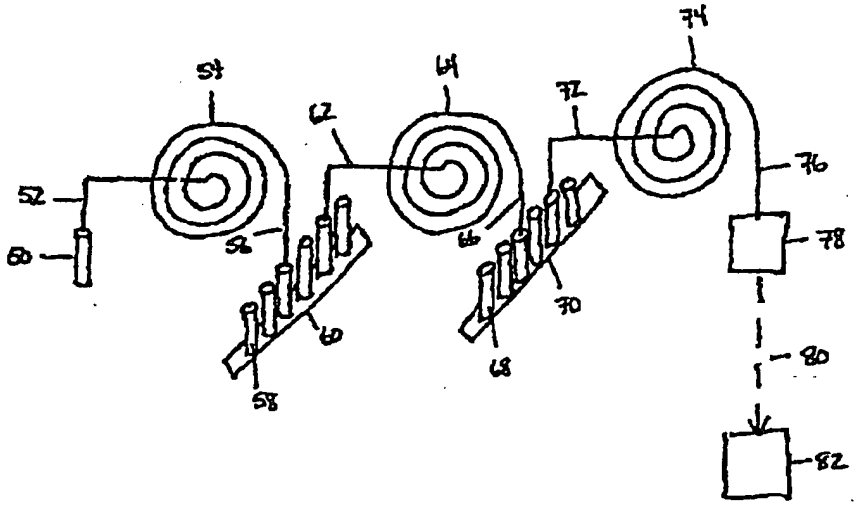
(71) Applicant *(for all designated States except US)*: TARGET DISCOVERY, INC. [US/US]; 1539 Industrial Road, San Carlos, CA 94070 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: SCHNEIDER, Luke, V. [US/US]; One Johnson Pier, C-30, Half Moon Bay, CA 94019 (US). HALL, Michael, P. [US/US]; 1364 Laural Street, #11, San Carlos, CA 94070 (US). PETESCH, Robert [US/US]; 6004 Robertson Avenue, Newark, CA 94560 (US). PETERSON, Jeffrey, N. [US/US]; 704 Bounty Drive, #410, Foster City, CA 94404 (US).

(74) Agents: KEZER, William, B. et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th floor, San Francisco, CA 94111 (US).

(54) Title: POLYPEPTIDE FINGERPRINTING METHODS, METABOLIC PROFILING, AND BIOINFORMATICS DATABASE



(57) Abstract

The invention provides methods, compositions, apparatus, and a computer data retrieval system for conducting proteomics and metabolic profiling on biological samples. One apparatus comprises: a sample container (50); a plurality of separation capillaries (54, 64, 74); a plurality of fraction collection devices (60, 70); a detector (78); and an analyzer (82).

POLYPEPTIDE FINGERPRINTING METHODS, METABOLIC PROFILING, AND BIOINFORMATICS DATABASE

CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims the benefit of U.S. provisional application 60/130,238, filed April 20, 1999. This application is also related to U.S. provisional application 60/075,715 filed February 24, 1998; copending U.S. patent application number 09/513,486, filed February 25, 2000, entitled "Protein Separation Via Multidimensional Electrophoresis," and having attorney docket number 020444-000200US; copending U.S. patent application number 09/513,395, filed February 25, 2000, entitled "Methods for Protein Sequencing," and having attorney docket number 020444-000300US; copending U.S. application number 09/513,907, filed February, 25, 2000, entitled "Polypeptide Fingerprinting Methods and Bioinformatics Database System," and having attorney docket number 020444-000100US; copending U.S. patent application number _____, filed April 19, 2000, entitled "Methods for Conducting Metabolic Analyses", and having attorney docket number 020444-000400US; and copending PCT application _____, filed April 19, 2000, entitled "Labeling of Protein Samples", and having attorney docket number 020444-000500. All of these applications are incorporated by reference in their entirety for all purposes.

FIELD OF THE INVENTION

This invention relates to the fields of protein separation and proteomics, metabolite profiling, bioinformatics, medicine, and computer databases.

BACKGROUND OF THE INVENTION

A goal of genomics research and differential gene expression analysis is to develop correlations between gene expression and particular cellular states (*e.g.*, disease states, particular developmental stages, states resulting from exposure to certain environmental stimuli and states associated with therapeutic treatments). Such correlations have the potential to provide significant insight into the mechanism of disease, cellular development and differentiation, as well as in the identification of new therapeutics, drug targets, and disease markers. Correlations of patterns of gene expression can also be used to provide similar insights into disease and organism

metabolism that can be used to speed the development of agricultural products, transgenic species, and for metabolic engineering of organisms to increase bioproduct yields or desirable metabolic activities.

Many functional genomic studies focus on changes in mRNA levels as being indicative of a cellular response to a particular condition or state. Recent research, however, has demonstrated that often there is a poor correlation between gene expression as measured by mRNA levels and actual active gene product formed (*i.e.*, protein encoded by the mRNA). [4] This finding is not surprising since many factors—including differences in translational efficiency, turnover rates, extracellular expression or compartmentalization, and post-translational modification—affect protein levels independently of transcriptional controls. Thus, the evidence indicates that functional genomics is best accomplished by measuring actual protein levels (*i.e.*, utilizing proteomic methods) rather than with nucleic acid based methods. The successful use of proteins for functional genomic analyses, however, requires reproducible quantification and identification of individual proteins expressed in cell or tissue samples.

It is at the protein level that metabolic control is exercised in cells and tissues. Comparison of the levels of protein expression between healthy and diseased tissues, or between pathogenic and nonpathogenic microbial strains, can speed the discovery and development of new drug compounds or agricultural products. Analysis of the protein expression pattern in diseased tissues or in tissues excised from organisms undergoing treatment can also serve as diagnostics of disease states or the efficacy of treatment strategies, as well as provide prognostic information regarding suitable treatment modalities and therapeutic options for individual patients.

Many proteins are expressed at varying levels in different cells. Proteins extracted from tissue or cell samples, using conventional techniques, must first be separated into individual proteins by gel or capillary electrophoresis or affinity techniques, before the individual proteins levels can be compared both within a sample and across samples obtained from different tissue sources. Because of the number of proteins expressed by a cell at any given time, multiple electrophoretic techniques (e.g., isoelectric focussing followed by electroporation through a polyacrylamide gel) are often applied to isolate all the individual proteins contained in a given sample.

Several techniques have been used to quantify the relative amounts of each protein present after the separation, including: staining proteins separated in a polyacrylamide gel with dyes (e.g., Brilliant Blue and Fast Green), with colloidial metals (e.g., gold or silver staining), or by prior labelling of the proteins during cellular synthesis by the addition of radioactive compounds (e.g., with 35S-methionine or 14C-amino acids, or 3H-leucine). Staining techniques yield poorly quantitative results because varying amounts of stain are incorporated into each protein and the stained protein must be resolved against the stained background of the gel or electroblotting substrate. Since radioactive labels are applied only to the proteins prior to separation, they overcome the background problem of staining techniques. However, feeding radioactive compounds to human subjects or handling radioactive materials in an uncontrolled field environment (e.g., crop plants) restricts the usefulness of this approach. Both staining and radiolabelling techniques also require inordinately long times to achieve detection. Staining and destaining of gels is a diffusion limited process requiring hours. Radiolabels must be quantified by exposing the labelled gel to photographic film or a phosphor screen for several hours to days while waiting for the radioactive decay process to produce a quantitative image. Direct infrared spectrophotometric interrogation of the proteins in a gel has also been used previously as a method for providing quantitative protein expression data. However, this quantitative resolution possible from this approach is adversely affected by variations in gel thickness and differential spreading of the protein spot between gels (changing the local concentration). Furthermore, the comparatively low absorption cross-section of proteins in the infrared limits the detection sensitivity. Analysis of the protein expression pattern does not provide sufficient information for many applications.

Several methods have also been proposed for the identification of proteins once they are resolved. The most common methods involve referencing the separation coordinates of individual proteins (e.g., isoelectric point and apparent molecular weight) to those obtained from archived separation coordinate data (e.g., annotated 2-D gel image databases) or control samples, performing a chemilytic or enzymatic digestion of a protein coupled with determination of the mass of the resulting peptide fragments and correlating this peptide mass fingerprint with that predicted to arrise from the predicted genetic sequence of a set of known proteins (*see* James, P., M. Quandroni, E. Carafoli,

and G. Gonnet, *Biochem. Biophys. Res. Commun.*, **195**:58-64 (1993); Yates, J.R., S. Speicher, P.R. Griffin, and T. Hunkapiller, *Anal. Biochem.*, **214**:397-408 (1993)), the generation of a partial protein sequence that is compared to the predicted sequences obtained from a genomic database (*see* Mann, M., paper presented at the IBC Proteomics conference, Boston, MA (Nov 10-11, 1997); Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweiger, T. Fotsis and M. Mann, *Nature*, **379**:466-469 (1996); Chait, B.T, R. Wang, R.C. Beavis and S.B.H. Kent, *Science*, **262**:89-92 (1993)), or combinations of these methods (*see* Mann, M., paper presented at the IBC Proteomics conference, Boston, MA (Nov 10-11, 1997); Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweiger, T. Fotsis and M. Mann, *Nature*, **379**:466-469 (1996); Chait, B.T, R. Wang, R.C. Beavis and S.B.H. Kent, *Science*, **262**:89-92 (1993)). Recent work indicates that proteins can only be unambiguously identified through the determination of a partial sequence, called a protein sequence tag (PST), that allows reference to the theoretical sequences determined from genomic databases (*see* Clauser, K.R., S. C. Hall, D. M. Smith, J.W. Webb, L.E. Andrews, H. M. Tran, L.B. Epstein, and A.L. Burlingame, " *Proc. Natl. Acad. Sci. (USA)*, **92**:5072-5076 (1995); Li, G., M. Walthan, N. L. Anderson, E. Unworth, A. Treston and J. N. Weinstein, *Electrophoresis*, **18**:391-402 (1997)). However, between 8 to 18 hours is currently required to generate a PST for a single protein sample by conventional techniques, with a substantial fraction of this time devoted to recovery of the protein sample from the separation method in a form suitable for subsequent sequencing (*see* Shevchenko, A., *et al.*, *Proc. Natl. Acad. Sci. (USA)*, **93**:14440-14445 (1996); Mark, J., paper presented at the PE/Sciex Seminar Series, Protein Characterization and Proteomics: Automated high throughput technologies for drug discovery, Foster City, CA (March, 1998). This makes the identification of all separated proteins from a tissue a time and cost prohibitive endeavour. This has restricted more widespread use of proteomic methods, despite their advantages for functional genomics and inhibited the development of proteomic databases, analogous to the genome databases now available (e.g., Genbank and the Genome Sequence Database).

Thus, current methods for identifying and quantitating the protein expression patterns ("protein fingerprints") of cells, tissues, and organs are lacking sufficient resolution, precision, and/or sensitivity. The present invention addresses these features lacking in the methods known in the art.

Polypeptide Separation Methods: Capillary Electrophoresis

Two-dimensional (2-D) gel electrophoresis is currently the most widely adopted method for separating individual proteins isolated from cell or tissue samples [5, 6, 7]. Evidence for this is seen in the proliferation (more than 20) of protein gel image databases, such as the Protein-Disease Database maintained by the NIH [8]. These databases provide images of reference 2-D gels to assist in the identification of proteins in gels prepared from various tissues.

Capillary electrophoresis (CE) is a different type of electrophoresis, and involves resolving components in a mixture within a capillary to which an electric field is applied. The capillary used to conduct electrophoresis is filled with an electrolyte and a sample introduced into one end of the capillary using various methods such as hydrodynamic pressure, electroosmotically-induced flow, and electrokinetic transport. The ends of the capillary are then placed in contact with an anode solution and a cathode solution and a voltage applied across the capillary. Positively charged ions are attracted towards the cathode, whereas negatively charged ions are attracted to the anode. Species with the highest mobility travel the fastest through the capillary matrix. However, the order of elution of each species, and even from which end of the capillary a species elutes, depends on its apparent mobility. Apparent mobility is the sum of a species electrophoretic mobility in the electrophoretic matrix and the mobility of the electrophoretic matrix itself relative to the capillary. The electrophoretic matrix may be mobilized by hydrodynamic pressure gradients across the capillary or by electroosmotically-induced flow (electrosomotic flow).

A number of different electrophoretic methods exist. Capillary isoelectric focusing (CIEF) involves separating analytes such as proteins within a pH gradient according to their isoelectric point (*i.e.*, the pH at which the analyte has no net charge) of the analytes. A second method, capillary zone electrophoresis (CZE) fractionates analytes on the basis of their intrinsic charge-to-mass ratio. Capillary gel electrophoresis (CGE) is designed to separate proteins according to their molecular weight. (For reviews of electrophoresis generally, and CIEF and CZE specifically, *see, e.g.*, Palmieri, R. and Nolan, J.A., "Protein Capillary Electrophoresis: Theoretical and Experimental Considerations for Methods Development," in *CRC Handbook of Capillary*

Electrophoresis: A Practical Approach, CRC Press, chapter 13, pp. 325-368 (1994) (electrophoresis generally); Kilar, F., "Isoelectric Focusing in Capillaries," in *CRC Handbook of Capillary Electrophoresis: A Practical Approach*, CRC Press, chapter 4, pp. 325-368 (1994); and McCormick, R.M., "Capillary Zone Electrophoresis of Peptides," in *CRC Handbook of Capillary Electrophoresis: A Practical Approach*, CRC Press, chapter 12, pp. 287-323 (1994). All of these references are incorporated by reference in their entirety for all purposes).

While 2-D gel electrophoresis is widely practiced, several limitations restrict its utility in functional genomics research. First, because 2-D gels are limited to spatial resolution, it is difficult to resolve the large number of proteins that are expressed in the average cell (1000 to 10,000 proteins). High abundance proteins can distort carrier ampholyte gradients in capillary isoelectric focusing electrophoresis and result in crowding in the gel matrix of size seiving electrophoretic methods (*e.g.*, the second dimension of 2-D gel electrophoresis and CGE), thus causing irreproducibility in the spatial pattern of resolved proteins [20, 21 and 22]. High abundance proteins can also precipitate in a gel and cause streaking of fractionated proteins [20]. Variations in the crosslinking density and electric field strength in cast gels can further distort the spatial pattern of resolved proteins [23, 24]. Another problem is the inability to resolve low abundance proteins neighboring high abundance proteins in a gel because of the high staining background and limited dynamic range of gel staining and imaging techniques [25, 22]. Limitations with staining also make it difficult to obtain reproducible and quantifiable protein concentration values. In some recent experiments, for example, investigators were only able to match 62% of test spots of the spots formed on 37 gels run under similar conditions [21; see also 28, 29]. Additionally, many proteins are not soluble in buffers compatible with acrylamide gels, or fail to enter the gel efficiently because of their high molecular weight [26, 27].

Thus, currently used methods of capillary electrophoresis provide significant limitations with regard to their usefulness is providing a detailed protein expression fingerprint of a cell or tissue sample.

Protein Species Identification/ Protein Sequence Tags

In contrast to characterizing proteins on the basis of their electrophoretic mobility or isoelectric point, an approach to identifying the protein species that are expressed in a tissue or cell sample is to obtain partial or complete peptide sequence information from proteins purified from the sample. Needless to say, but this approach is laborious and is of limited sensitivity as it requires extensive and often problematic purification steps to isolate individual protein species to allow for unambiguous sequence determination, and in many cases is simply not feasible for proteins which are not highly abundant and/or are not readily purifiable free from contaminant protein species.

It is also important that primary amino acid sequence or a partial sequence (i.e., a protein sequence tag, "PST") be determined so that the reason underlying changes in the protein expression pattern related to proteins that appearing at different separation coordinates, can be determined. Proteins may appear at more than one separation coordinate, depending on the degree of post-translational modification exercised on that protein by the cell or tissue. The separation coordinate for a protein may also change due to genetic mutations. Changes in the relative abundance of a protein at any given separation coordinate may also be due to changes in the regulation of gene expression. Only by unambiguously identifying each of the proteins resolved can the reason underlying any variations in protein expression across different samples be deduced.

Several methods have previously been proposed for determining the sequence or a protein sequence tag of separated proteins. These include: sequential rounds of N-terminal or C-terminal labeling followed by liberation and determination of the labeled amino acid, exoproteolytic digestion of the protein one amino acid at a time, endoproteolytic digestion of larger proteins into smaller peptides followed by N- and C-terminal labeling and amino acid determination, and mass spectrometric fragmentation pattern recognition. Sequential labeling and digestion techniques (e.g., Edman chemistry) are time consuming, even when automated, because the process must be repeated through many cycles before a sufficiently large protein sequence tag can be accumulated. Propagation of errors-i.e., either from incomplete labeling on each round, incomplete liberation of the labeled amino acid, or both-also limits the length of protein sequence that can be determined using these techniques. While a more complete protein sequence can be obtained by first using endoproteases to cleave the protein into smaller fragments prior

to application of the sequential labeling and digestion chemistry, this also introduces the time and labor intensive step of reseparating and purifying the protein fragments, usually by reapplication of an electrophoretic separation technique. Determining the sequence order of these peptide fragments in the original protein can also present additional problems. Carboxy-terminal methoxy labeling of cyanogen bromide digests has been used to identify the C-terminal peptide fragment from other fragments formed by cyanogen bromide digestion of a larger protein.

Protein Sequence Determination by Mass Spectrometry

Mass spectrometric techniques are increasingly being applied to protein identification because of their speed advantage over the more traditional methods. Electrospray and matrix assisted laser desorption ionization (MALDI) are the most common mass spectrometric techniques applied to protein analysis because they are best able to ionize large, low volatility, molecular species. Two basic strategies have been proposed for the MS identification of proteins after separation: 1) mass profile fingerprinting ('MS fingerprinting')and 2) sequencing of one or more peptide domains by MS/MS ('MS/MS sequencing'). MS fingerprinting is achieved by accurately measuring the masses of several peptides generated by a proteolytic digest of the intact protein and searching a database for a known protein with that peptide mass fingerprint. MS/MS sequencing involves actual determination of one or more PSTs of peptides derived from the protein digest by generation of sequence-specific fragmentation ions in the quadrapole of an MS/MS instrument. Refinements in both of these techniques have also reduced the amount of individual proteins needed to achieve signature detection.

In one approach, a protein is chemilytically (e.g., cyanogen bromide) or enzymatically (e.g., trypsin) digested at sequence specific sites to form peptides. The specificity of the cleavage yields peptides of reproducible masses that can subsequently be determined by MS. The mass spectrometric peptide pattern detected from an individual protein is then compared to a database of similar patterns generated from purified proteins with known sequences or predicted from the theoretical protein sequence based on the expected digestion pattern. The identity of the unknown protein is then inferred to be that of the known protein that best matches its peptide mass fingerprint.

Historically, techniques such as Edman degradation have been extensively used for protein sequencing. However, sequencing by collision-induced dissociation MS methods (MS/MS sequencing) has rapidly evolved and has proved to be faster and require less protein than Edman techniques. MS sequencing is accomplished either by using higher voltages in the ionization zone of the MS to randomly fragment a single peptide isolated from a protein digest, or more typically by tandem MS using collision-induced dissociation in the ion trap (quadrapole). However, the application of CID methods to protein sequencing require that the protein first be chemilytically or enzymatically digested.

Several techniques can be used to select the peptide fragment used for MS/MS sequencing, including accumulation of the parent peptide fragment ion in the quadrapole MS unit, capillary electrophoretic separation coupled to ES-TOF MS detection, or other liquid chromatographic separations. The amino acid sequence of the peptide is deduced from the molecular weight differences observed in the resulting MS fragmentation pattern of the peptide using the published masses associated with individual amino acid residues in the MS, and has been codified into a semi-autonomous peptide sequencing algorithm. In this approach the peptide to be sequenced is typically accumulated in the quadrapole of a mass spectrometer. CID is then accomplished by injecting a neutral collision gas, typically Ar, into this ion trap to force high energy collisions with the peptide ion. Some of these collisions result in cleavage of the peptide backbone and the generation of smaller ions that, by virtue of their different mass to charge ratio, leave the quadrapole and are detected. The majority of the peptide cleavage reactions occur in a relatively few number of ways, resulting in a high abundance of certain types of cleavage ions. The peptide sequence is then deduced from the apparent masses of these high abundance peptide fragments detected.

Mass spectrometry has the additional advantage in that it can be efficiently coupled to electrophoretic separation techniques both with or without endoproteolytic (e.g., trypsin digestion) or chemilytic (e.g., cyanogen bromide) cleavage of the protein into smaller fragments. However, no mass spectrometric technique has previously been described that directly determines the protein sequence or a protein sequence tag of unknown proteins. Furthermore, no MS sequencing technique has previously been

described that directly couples to electrophoretic methods used to separate large numbers of proteins from a mixed protein sample.

For example, in the mass spectrum of a 1425.7 Da peptide (HSDAVFTDNYTR) isolated in an MS/MS experiment acquired in positive ion mode, the difference between the full peptide 1425.7 Da and the next largest mass fragment (y_{11}, 1288.7 Da) is 137 Da. This corresponds to the expected mass of an N-terminal histidine residue that is cleaved at the amide bond. For this peptide, complete sequencing is possible as a result of the generation of high-abundance fragment ions that correspond to cleavage of the peptide at almost every residue along the peptide backbone. The generation of an essentially complete set of positively-charged fragment ions that include either end of the peptide is a result of the basicity of both the N- and C-terminal residues (H and R, respectively). If a basic residue is located at the N- or C-terminus, especially R, most of the ions produced in the CID spectrum will contain that residue since positive charge is essentially localized at that site. This greatly simplifies the resulting spectrum since these basic sites direct the fragmentation into a limited series of specific daughter ions. Peptides that lack basic residues tend to fragment into a more complex mixture of fragment ions that makes sequence determination more difficult.

Extending this idea, others demonstrated that attaching a hard positive charge to the N-terminus is an effective approach for directing the production of a complete series of N-terminal fragment ions from a parent peptide in CID experiments regardless of the presence of a basic residue at the N-terminus. Theoretically, all fragment ions are produced by charge-remote fragmentation directed by the fixed-charged group. Peptides have now been modified with several classes of fixed-charged groups, including dimethylalkylammonium, substituted pyridinium, quaternary phosphonium, and sulfonium derivatives. The characteristics of the most desirable labels are that they are easily synthesized, increase the ionization efficiency of the peptide, and (most importantly) direct the formation of a specific fragment ion series with minimal unfavorable label fragmentation. The most favorable derivatives that satisfy these criteria are those of the dimethylalkylammonium class with quarternary phosphonium derivatives being only less favorable due to their more difficult synthesis. Substituted pyridinium derivatives are better suited for high-energy CID as opposed to alkylammonium derivatives.

Despite some progress in peptide analysis, protein identification remains a major bottleneck in field of Proteomics, with up to 18 hours being required to generate a protein sequence tag of sufficient length to allow the identification of a single purified protein from its predicted genomic sequence. Unambiguous protein identification is attained by generating a protein sequence tag (PST), which is now preferentially accomplished by collision-induced dissociation in the quadrapole of an MS/MS instrument. Limitations on the ionization efficiency of larger peptides and proteins restrict the intrinsic detection sensitivity of MS techniques and inhibit the use of MS for the identification of low abundance proteins. Limitations on the mass accuracy of time of flight (TOF) detectors can also constrain the usefulness of MS/MS sequencing, requiring that proteins be digested by proteolytic and chemolytic means into more manageable peptides prior to sequencing. Clearly, rapid and cost effective protein sequencing techniques would improve the speed and lower the cost of proteomics research. Finally, the separation agents and buffers used in traditional protein separation techniques are often incompatible with MS identification methods.

Labeling of Protein Samples

The correlation of protein expression levels obtained from healthy and diseased tissue is the basis of proteomics research. Proteins extracted from tissue or cell samples typically must be separated into individual proteins by gel electrophoresis (O'Farrel, P.H., *J Biol. Chem.*, **250**:4007 (1975); Hochstrasser, D.F., et al., *Anal Biochem.*, **173**:424 (1988); Hühmer, A. F. R., et al., *Anal. Chem.*, **69**:29R-57R (1997); Garfin, D.E., *Methods in Enzymology*, **182**:425 (1990)), capillary electrophoresis (Smith, R. D., et al., "Capillary electrophoresis-mass spectrometry," in: *CRC Handbook of Capillary Electrophoresis: A Practical Approach*, Chp. 8, pgs 185-206 (CRC Press, Boca Raton, FL, 1994); Kilár, F., "Isoelectric focusing in capillaries," in: *CRC Handbook of Capillary Electrophoresis: A Practical Approach*, Chp. 4, pgs. 95-109 (CRC Press, Boca Raton, FL, 1994); McCormick, R. M., "Capillary zone electrophoresis of peptides," in: *CRC Handbook of Capillary Electrophoresis: A Practical Approach*, Chp. 12, pgs 287-323 (CRC Press, Boca Raton, FL, 1994); Palmieri, R. and Nolan, J. A., "Protein capillary electrophoresis: theoretical and experimental considerations for methods development," in: *CRC Handbook of Capillary Electrophoresis: A Practical Approach*, Chp. 13, pgs

WHAT IS CLAIMED IS:

1. A method for separating a polypeptide species from a sample solution containing a plurality of polypeptide species and identifying said polypeptide species, the method comprising electrophoresing said sample solution containing a plurality of polypeptide species in a capillary electrophoresis device to separate and elute polypeptide species thereby resolving said protein species based on at least one first biophysical parameter which discriminates protein species; and obtaining, by mass spectrographic fragmentation of eluted polypeptide species, a polypeptide sequence tag ("PST") identifying at least one resolved protein species.

2. The method of claim 1, wherein the method further comprises electrophoresing polypeptide species eluted from said capillary electrophoresis device in a second capillary electrophoresis device to separate and elute polypeptide species thereby resolving said protein species based on at least one second biophysical parameter which discriminates protein species, prior to performing mass spectrographic fragmentation on the polypeptide species thereby obtained.

3. The method of claim 1, wherein the capillary electrophoresis device is a capillary isoelectric focusing (CIEF) device, a capillary zone electrophoresis device (CZE), or a capillary gel electrophoresis device (CGE).

4. The method of claim 2, wherein the capillary electrophoresis device is a CIEF device and the second capillary electrophoresis device is either a CZE device or a CGE device.

5. The method of claim 2, wherein the capillary electrophoresis device is a CZE device and the second capillary electrophoresis device is either a CIEF device or a CGE device.

6. The method of claim 2, wherein the capillary electrophoresis device is a CGE device and the second capillary electrophoresis device is either a CIEF device or a CZE device.

7. The method of claim 2, wherein the method further comprises electrophoresing polypeptide species eluted from said second capillary electrophoresis device in a thirdcapillary electrophoresis device to separate and elute polypeptide species thereby resolving said protein species based on at least one third biophysical parameter which discriminates protein species, prior to performing mass spectrographic fragmentation on the polypeptide species thereby obtained.

8. The method of claim 7, wherein the the capillary electrophoresis device is a CIEF device and the second capillary electrophoresis device is either a CZE device and the third capillary electrophoresis device is a CGE device.

9. The method of claim 1, wherein the sample solution containing a plurality of polypeptide species comprises labeled polypeptide species.

10. The method of claims 1, 2, and 7 wherein the polypeptide species are labeled after capillary electrophoresis and prior to mass spectroscopy.

11. The method of claims 9 and 10, wherein the label comprises a detectable moiety.

12. The method of claims 9 and 10, wherein the label comprises an ion mass signature component.

13. The method of claims 9 and 10, wherein the label comprises an ion mass signature component and a detectable moiety.

14. A method for identifying a high-resolution protein expression fingerprint for a cell type, tissue, or pathological sample, comprising obtaining a protein-containing extract of a cellular sample and electrophoresing said extract with a first capillary electrophoresis

apparatus, eluting protein-containing fractions therefrom, electrophoresing said protein containing fractions on a second capillary electrophoresis apparatus, or plurality thereof in parallel, and identifying the species of proteins by fragmentation mass spectroscopy sequencing to obtain PSTs for a plurality of protein species, and compiling a dataset or fingerprint record containing the collection of PSTs obtained thereby.

15. The method of claim 14, comprising quantitative detection of protein species and compiling a dataset wherein the relative abundance and/or absolute amount of a plurality of protein species eluted from said second capillary electrophoresis are cross-tabulated with the PST identification.

16. A computer system comprising: a database including a plurality of fingerprint records each comprising an array of at least 50 molecular species each having a unique identifier cross-tabulated with quantitative data indicating relative and/or absolute abundance of each species in a sample, and a user interface capable of receiving a selection of one or more queries to said database for use in determining a rank-ordered similarity of fingerprint records in the database.

17. A computer system of claim 16 having a fingerprint record comprising an array of at least 50 protein species each having a PST cross-tabulated with a separation coordinate produced by the method of claim 1.

18. A computer system of claim 16 having a fingerprint record comprising an array of at least 50 protein species each having a PST obtained by the method of claim 1.

19. A method for producing or accessing a computer database comprising a computer and software for storing in computer-retrievable form a collection of protein expression fingerprint records cross-tabulated with data specifying the source of the protein-containing sample from which each protein expression fingerprint record was obtained.

20. The method of claim 19, wherein at least one of the sources is from a tissue sample known to be free of pathological disorders.

21. The method of claim 19, wherein at least one of the sources is a known pathological tissue specimen.

22. A method of labeling a plurality of different proteins in a protein sample, said method comprising contacting said protein sample with a labeling agent comprising a unique ion mass signature component, a quantitative detection component and a reactive functional group to covalently attach a label to at least a portion of said plurality of different proteins.

23. A method in accordance with claim 22, wherein said protein sample comprises at least five different proteins.

24. A method in accordance with claim 22, wherein said detection enhancement component is a fluorophore selected from the group consisting of naphthylamines, coumarins, acridines, stilbenes and pyrenes.

25. A method in accordance with claim 22, wherein said detection enhancement component is a fluorophore selected from the group consisting of 1-dimethylaminonaphthyl-5-sulfonate, 1-anilino-8-naphthalene sulfonate, 2-p-toluidinyl-6-naphthalene sulfonate, 3-phenyl-7-isocyanatocoumarin, 9-isothiocyanatoacridine, acridine orange, N-(p-(2-benzoxazolyl)pheny)maleimide, and benzoxadiazoles.

26. A method for separating a plurality of proteins in an initial sample, comprising

performing a plurality of electrophoretic methods in series, each method comprising

electrophoresing a sample containing multiple proteins, whereby a plurality of resolved proteins are obtained, and

wherein the sample electrophoresed contains only a subset of the plurality of resolved proteins from the immediately preceding method in the series, except the first method of the series in which the sample is the initial sample; and

detecting resolved proteins from the final electrophoretic method.

27. The method of claim 26, wherein the plurality of electrophoretic methods are capillary electrophoresis methods.

28. The method of claim 27, wherein the plurality of capillary electrophoretic methods are selected from the group consisting of capillary isoelectric focusing electrophoresis, capillary zone electrophoresis and capillary gel electrophoresis.

29. The method of claim 27, wherein the plurality of capillary electrophoretic methods are two methods, the first electrophoretic method being capillary isoelectric focusing electrophoresis and the second electrophoretic method being capillary gel electrophoresis.

30. The method of claim 27, wherein the plurality of capillary electrophoretic methods are two methods, the first electrophoretic method being capillary zone electrophoresis and the second electrophoretic method being capillary gel electrophoresis.

31. The method of claim 27, wherein the plurality of capillary electrophoretic methods are three methods, the first, second and third electrophoretic methods being capillary isoelectric focusing electrophoresis, capillary zone electrophoresis and capillary gel electrophoresis, respectively.

32. The method of claim 26, wherein

the performing further comprises repeating the electrophoresing step multiple times, each time with a different sample containing only a subset of the plurality of resolved proteins from the immediately preceding method in the series, whereby a plurality of resolved proteins for each of the different samples is obtained; and

the detecting comprises detecting resolved proteins from each of the different samples from the final electrophoretic method.

33. A method for separating a plurality of proteins, comprising

performing a plurality of electrophoretic methods in series, wherein the method or methods preceding the final method comprise

withdrawing and collecting multiple fractions containing proteins resolved during the electrophoretic method, and

wherein each electrophoretic method is conducted with a sample from a fraction collected in the preceding electrophoretic method, except the first electrophoretic method which is conducted with a sample containing the plurality of proteins;

labeling the plurality of proteins or labeling protein contained in collected fractions prior to conducting the last electrophoretic method; and

detecting protein contained in electrophoretic medium utilized during a final electrophoretic method by detecting label borne by the protein, the final electrophoretic method being performed with a sample from one or more fractions obtained in the penultimate electrophoretic method.

34. The method of claim 33, wherein the detecting step comprises detecting protein with a detector in fluid communication with a separation cavity containing the electrophoretic medium utilized during the final electrophoretic method.

35. The method of claim 33, wherein one of the plurality of electrophoretic methods is capillary zone electrophoresis and the labeling step is conducted prior to conducting the capillary zone electrophoresis method.

36. The method of claim 33, wherein one of the plurality of electrophoretic methods is capillary isoelectric focusing and the labeling step is performed subsequent to the capillary isoelectric focusing method.

37. A method for separating a plurality of proteins, comprising

performing one or more capillary electrophoretic methods, each of the one or more methods comprising

electrophoresing a sample containing multiple proteins within an electrophoretic medium contained within a capillary; and

withdrawing and collecting multiple fractions, each fraction containing proteins resolved during the electrophoresing step, and

wherein each method is conducted with a sample from a fraction collected in the preceding electrophoretic method, except the first electrophoretic method which is conducted with a sample containing the plurality of proteins;

labeling the plurality of proteins or labeling protein contained in collected fractions prior to conducting the last electrophoretic method; and

conducting a final capillary electrophoresis method with a final capillary, the final method comprising detecting resolved protein within, or withdrawn from, the final capillary.

38. The method of claim 37, wherein the one or more capillary electrophoresis methods is capillary isoelectric focusing electrophoresis and the final capillary electrophoresis method is capillary gel electrophoresis.

39. The method of claim 37, wherein the one or more capillary electrophoresis methods is capillary zone electrophoresis and the final capillary electrophoresis method is capillary gel electrophoresis.

40. The method of claim 37, wherein the one or more capillary electrophoresis methods is two methods, the first method being capillary isoelectric focusing and the second method being capillary zone electrophoresis, and the final capillary electrophoresis method is capillary gel electrophoresis.

41. A method for separating a plurality of proteins in an initial sample, comprising:

performing a plurality of electrophoretic methods in series, each method comprising

electrophoresing within an electrophoretic medium a sample containing multiple proteins whereby fractions containing a subset of the multiple proteins are isolated physically, temporally or spatially, and

wherein the sample electrophoresed is obtained from a fraction isolated during the immediately preceding method in the series, except the first method of the series in which the sample is the initial sample; and

detecting isolated proteins from the final electrophoretic method.

42. A method for separating a plurality of proteins, comprising performing at least two capillary electrophoretic separations in series, wherein a sample for the second capillary electrophoretic separation is from a fraction obtained during the first capillary electrophoretic separation, the fraction containing only a subset of the plurality of proteins contained in the sample electrophoresed during the first capillary electrophoretic method.

43. A method for analyzing metabolic pathways, comprising:

administering to a subject a substrate labeled with a stable isotope, wherein the relative isotopic abundance of the isotope in the substrate is known;

allowing the labeled substrate to be at least partially metabolized by the subject to form one or more target metabolites; and

determining the abundance of the isotope in a plurality of target analytes in a sample from the subject to determine a value for the flux of each target analyte, wherein the plurality of target analytes comprise the substrate and/or one or more of the target metabolites.

44. The method of claim 43, wherein the determining comprises at least partially separating the target analytes from other biological components in the sample prior to determining the flux values.

45. The method of claim 44, wherein the separating comprises performing a plurality of capillary electrophoresis methods in series.

46. The method of claim 45, wherein the performing of the capillary electrophoresis methods generate separate fractions for at least one class of metabolite, wherein the class of metabolite is selected from the group consisting of proteins, polysaccharides, carbohydrates, nucleic acids, amino acids, nucleotides, nucleosides, fats, fatty acids and organic acids.

47. The method of claim 43, wherein the determining comprises obtaining multiple samples from the subject at different predetermined time points, separating the target analytes from other biological components in each of the samples, and determining the abundance of the isotope in the target analytes contained in each sample, whereby a plurality of values for the abundance of the isotope in each target analyte are obtained, the flux value for each target analyte being determined from the plurality of abundance values determined for it.

48. A method for analyzing metabolic pathways, comprising:

separating at least partially a plurality of target analytes from biological components contained in a sample obtained from a subject, the target analytes comprising a substrate labeled with a stable isotope and/or one or more target metabolites resulting from the metabolism of the substrate by the subject, and wherein the relative isotopic abundance of the isotope in the substrate is known; and

determining the abundance of the isotope in a plurality of the target analytes in the sample to determine a value for the flux of each target analyte.

49. A method for screening for metabolites correlated with a disease, comprising:

analyzing a sample from a test subject having the disease, the sample comprising a substrate labeled with a stable isotope administered to the test subject and/or one or more target metabolites resulting from metabolism of the substrate by the test subject, the relative isotopic abundance of the isotope in the substrate known at the time of administration, and wherein the analyzing step comprises determining the isotopic abundance of the isotope in a plurality of analytes in the sample to determine a value for the flux of each analyte, wherein the plurality of analytes comprise the substrate and/or one or more of the target metabolites; and

comparing flux values for the analytes with flux values for the same analytes obtained for a control subject, wherein a difference in a flux value for an analyte indicates that such analyte is correlated with the disease.

50. A method for screening for the presence of a disease, comprising:

analyzing a sample from a test subject, the sample comprising a substrate labeled with a stable isotope administered to the test subject and/or one or more target metabolites resulting from metabolism of the substrate by the test subject, the relative isotopic abundance of the isotope in the substrate known at the time of administration, and wherein the analyzing step comprises determining the abundance of the isotope in a plurality of analytes in the sample to determine a value for the flux of each analyte, wherein the plurality of analytes comprise the substrate and/or one or more of the target metabolites; and

for each target analyte, comparing the determined flux value with a range of flux values for that analyte, wherein the range is known to be correlated with the disease and if a determined flux value for a target analyte falls within the range for that target analyte, it indicates that the test subject has the disease or is susceptible to the disease.

51. An apparatus for performing a method of claims 1, 14, 22, 26, 33, 37, 41, 43, 48, 49, or 50, comprising:

at least two capillary electrophoresis devices fixed to a common platform or frame and in liquid communication with each other and with a mass spectrometer wherein a sample flows into a first capillary electrophoresis device and separation of analytes occurs based on at least a first biochemical separation parameter and the analytes subsequently flow into a second capillary electrophoresis device and separation of analytes occurs based on at least a second biochemical separation parameter which is different than said first biochemical separation parameter and wherein the analytes subsequently flow into a mass spectrometer.

52. The use of a method of claims claims 1, 14, 19, 22, 26, 33, 37, 41, 43, 48, 49, or 50, or the apparatus of claim 51, or of the computer system or database of claim 16.